References

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.112. URL https://aclanthology.org/2020.emnlp-main.112.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.459. URL https://aclanthology.org/2021.acl-long.459.

Baoyu Jing, Zeya Wang, and Eric Xing. Show, describe and conclude: On exploiting the structure information of chest X-ray reports. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6570–6580, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1657. URL https://aclanthology.org/P19-1657.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL http://arxiv.org/abs/1412.6980.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13753–13762, June 2021a. URL https://openaccess.thecvf.com/content/CVPR2021/html/Liu_Exploring_and_Distilling_Posterior_and_Prior_Knowledge_for_Radiology_Report_CVPR_2021_paper.html.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest X-ray report generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 269–280, Online, August 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.23. URL https://aclanthology.org/2021.findings-acl.23.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, volume 32, pages 8024–8035. Curran Associates, 2019. URL https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

Appendix A. Preprocessing Details

Following the preprocessing of R2Gen, the raw documents are converted to lowercase and tokenized using the NLTK library. Furthermore, we removed redundant spaces, empty lines, serial numbers, and punctuation marks from the documents In IU X-RAY, we apply the widely-used splits in (Chen et al., 2020; Jing et al., 2019; Liu et al., 2021a) and partition the dataset into train/validation/test set by 7:1:2. Following the works (Chen et al., 2020, 2021), we remove the tokens whose frequency of occurrence in the training set is less than 3, resulting in 789 words for the entire dataset. In MIMIC, we adopt the official splits for the dataset to report our results: 368,960 for training, 2,991 for validation, and 5,159 for test. Following the works (Chen et al., 2020, 2021; Liu et al., 2021b), we keep the tokens with a frequency in the training set are more than 10.

Table 1: Summary of TIMER's performance improvements over baselines. $\hat{\Delta}$ indicates percentage improvements over baselines.

Methods (%)	BLEU_1	BLEU_2	BLEU_3	BLEU_4	Meteor	Rouge_L	Clinical Metric
IU X-RAY							
$\hat{\Delta}$	6.42	7.57	9.07	13.06	4.45	4.55	61.16
$\hat{\Delta}$-BiLSTM	17.96	10.89	12.10	20.14	8.72	11.64	45.09
$\hat{\Delta}$-R2GEN	1.11	1.75	2.58	5.02	0.84	3.10	48.38
$\hat{\Delta}$-CMN	8.37	10.13	11.04	12.58	7.32	4.00	45.61
$\hat{\Delta}$-CMM+RL	0.08	8.01	11.14	15.59	1.39	0.13	131.42
MIMIC							
$\hat{\Delta}$	11.27	9.66	9.01	10.50	8.64	3.54	49.30
$\hat{\Delta}$-BiLSTM	42.86	42.61	44.27	48.57	30.55	7.69	53.25
$\hat{\Delta}$-R2GEN	8.13	2.27	0.69	0.97	6.91	2.79	118.17
$\hat{\Delta}$-CMN	7.58	5.04	3.77	4.94	3.67	3.17	84.08
$\hat{\Delta}$-CMM+RL	0.52	1.76	1.03	3.79	1.17	1.23	167.48

Appendix B. Implementation Details

We use ADAM (Kingma and Ba, 2015) optimizer to train our model with the learning rate 0.001 and decay such rate by a factor of 0.9 per epoch for each dataset. We update the TIMER for each inner loop training in the IU X-RAY dataset and every 100 iterations of inner loop training in the MIMIC dataset. The max training epoch is 100 for the IU X-RAY and 30 for the MIMIC, due to the data sizes and our computational resources. We generate tokens by beam search (Sutskever et al., 2014) with 3 beam size in the test stage for all experiments. All implementations are on PyTorch (Paszke et al., 2019).

In baseline BiLSTM, we set the number of tags for semantic attention as 10 and all hidden states and word embeddings as 512. The learning rates for the CNN and the hierarchical LSTM are 1e-5 and 5e-4 respectively.

In baseline R2GEN, We adopt the ResNet101 to extract images' features with the dimension of each feature set to 2048. The dimension of relational memory in multi-head attention is 512 and contains 8 heads. The number of memory slots is set to 3 by default. The learning rate is 5e-5 for the visual extractor and 1e-4 for other parameters. We decay such rate by a factor of 0.8 per epoch for each dataset.

In baseline CMN, the image feature extractor has the same setting as R2GE. The encoder-decoder structure adopts Transformer with 3 layers and 8 attention heads. The memory matrix dimension is 512 and the number of memory vectors is set to 2048. The learning rates of the visual extractor and other parameters are set to 5×10^{-5} and 10^{-4}, respectively, and we decay them by a 0.8 rate per epoch for all datasets.

CMM+RL baseline keeps all the same settings as CMN. Following the setting in the paper, we adopt the greedy sampling method to generate reports for self-critical learning. We set the batch size as 8 since this achieves the best result in the paper's report.

Table 1 presents the details of our model improvements percentage over baselines.